Peningkatan Akurasi Klasifikasi Kualitas Udara melalui Oversampling dengan Metode Support Vector Machine dan Random Forest

I Gusti Ayu Nandia Lestari[1], Komang Agus Ady Aryanto[2]
[1]Institut Teknologi dan Bisnis STIKOM Bali, [2] Rajamangala University of Technology Thanyaburi
e-mail: [1]nandia@stikom-bali.ac.id, [2]i_komang@mail.rmutt.ac.th

Diajukan: 24 Mei 2023; Direvisi: 20 Juni 2023; Diterima: 21 Juni 2023

Abstrak

Kualitas udara merupakan faktor penting yang memengaruhi kesehatan manusia dan semua makhluk hidup. Penelitian ini bertujuan untuk mengembangkan model klasifikasi kualitas udara menggunakan teknik oversampling SMOTE, serta metode Random Forest dan Support Vector Machine. Data yang digunakan adalah Indeks Standar Pencemaran Udara (ISPU) DKI Jakarta selama tahun 2022. Hasil pengujian menunjukkan bahwa menggunakan Random Forest, akurasi model mencapai 98%, sedangkan dengan penerapan SMOTE, akurasi meningkat menjadi 99%. Pada pemodelan dengan Support Vector Machine, akurasi mencapai 91%, namun dengan SMOTE, akurasi meningkat menjadi 95%. Hal ini menunjukkan bahwa penggunaan teknik oversampling SMOTE dapat meningkatkan akurasi model. Diharapkan penelitian ini dapat memberikan kontribusi penting bagi pemantauan dan pengelolaan lingkungan, serta memberikan pemahaman yang lebih baik tentang kualitas udara.

Kata kunci: *Kualitas Udara, Support Vector Machine, Random Forest, SMOTE.*

Abstract

Air quality is an important factor that affects the health of humans and all living creatures. This study aims to develop a model for classifying air quality using the SMOTE oversampling technique, as well as the Random Forest and Support Vector Machine methods. The data used is the Air Pollution Standard Index (ISPU) of DKI Jakarta for the year 2022. The test results show that using Random Forest, the model accuracy reaches 98%, while with the application of SMOTE, the accuracy increases to 99%. In modeling with Support Vector Machine, the accuracy reaches 91%, but with SMOTE, the accuracy increases to 95%. This indicates that the use of SMOTE oversampling technique can improve model accuracy. It is hoped that this research will provide valuable contributions to environmental monitoring and management, as well as a better understanding of air quality.

Keywords*: Kualitas Udara, Support Vector Machine, Random Forest, SMOTE.*

1. Pendahuluan

Kualitas udara sangat berperan dalam mempengaruhi kesehatan dan kesejahteraan manusia. Polusi udara dapat mengakibatkan berbagai penyakit seperti gangguan pernapasan, masalah jantung, dan risiko kanker. Jakarta, sebagai Ibu Kota Indonesia, mengalami tingkat polusi udara yang tinggi, yang secara langsung mempengaruhi kualitas udara di wilayah tersebut. Informasi mengenai kualitas udara kepada Masyarakat yaitu dengan menggunakan Indeks Standar Pencemar Udara (ISPU). ISPU merupakan suatu angka tanpa satuan yang mengggambarkan kondisi kualitas udara ambien pada suatu lokasi tertentu. ISPU dihitung berdasarkan konsentrasi beberapa parameter pencemar udara seperti, partikulat (PM10), sulfur dioksida (SO2), nitrogen dioksida (NO2), karbon monoksida (CO), dan ozon (O3). Partikulat (PM10) merupakan partikulat udara dengan ukuran yang lebih kecil dari 10 mikrometer. Partikulat ini dapat masuk ke dalam saluran pernapasan dan menyebabkan masalah kesehatan seperti gangguan pernapasan dan kardiovaskular, Sulfur dioksida (SO2) merupakan gas beracun yang dihasilkan dari pembakaran bajan bakar fosil seperti batu bara dan minyak bumi. SO2 dapat menyebabkan iritasi pada saluran pernapasan dan dapat menjadi prekusor bagi pembentukan hujan asam. Nitrogen dioksida (NO2) merupakan Gas yang dihasilkan dari aktivitas manusia seperti kendaraan bermotor dan pembakaran bahan bakar. NO2 dapat menyebabkan iritasi pada paru-paru, meningkatkan risiko asma, dan mempengaruhi kulitas udara di

perkotaan. Karbon monoksida (CO) merupakan gas tak berwarna dan tidak berbau yang dihasilkan dari pembakaran bahan organik. Paparan CO dapat menyebabkan keracunan yang berbahaya karena dapat mengganggu transportasi oksigen dalam darah. Ozon (O3) merupakan komponen utama dari polusi ozon troposferik, yang terbentuk dari reaksi antara nitrogen dioksida (NO2) dan senyawa organic yang dihasilkan dari aktivitas manusia. Ozon dapat menyebabkan iritasi pada saluran pernapasan, memperburuk asma, dan mempengaruhi kesehatan manusia dan ekosistem. Dengan menggunakan teknologi machine learning, dapat digunakan untuk klasifikasi tingkat pencemaran berdasarkan konsentrasi polutan tertentu [1].

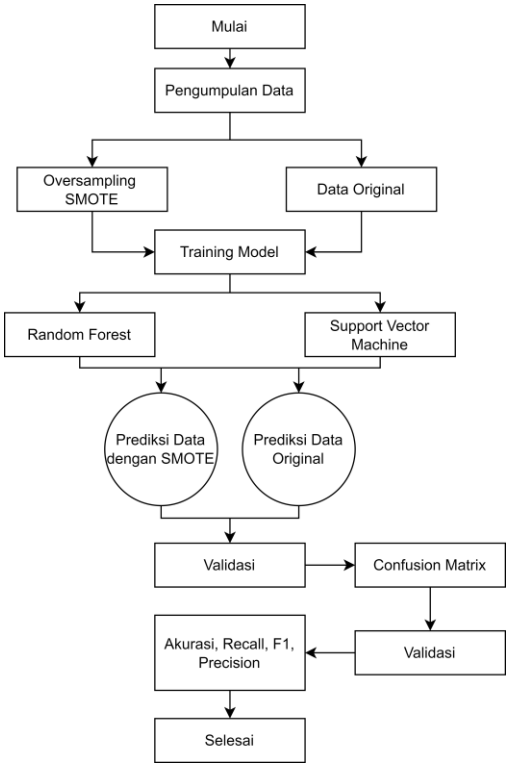
Dalam mengetahui pencemaran udara, diperlukan metode yang dapat memantau kualitas udara dan mengklasifikasikan kondisinya dengan memanfaatkan machine learning [1]. Machine Learning memiliki kemampuan mengklasifikasikan berbagai aspek seperti kualitas udara. Dengan mengidentifikasi tingkat kualitas udara dalam kategori baik, sedang, atau tidak sehat berdasarkan data jenis polutan seperti PM2.5, PM10, SO2, CO, O3, dan NO2 dalam udara [1]. Pada machine learning terdapat permasalahan yang sering juga terjadi dalam ketidakseimbangan data atau sering dinamakan imbalanced dataset penelitian [2]. Imbalanced dataset ini seringkali memberikan hasil yang tidak akurat sehingga perlu adanya imbalaced dataset dengan menggunakan machine learning. Dengan mengatasi permasalahan imbalaced dataset dengan melakukan oversampling. Teknik oversampling dengan Syntetic Minority Oversampling (SMOTE) dengan Edited Nearest Neighbors (ENN) dan TomekLinks terhadap Support Vector Machine (SVM). Hasil uji coba menunjukan bahwa teknik kombinasi ini efektif dalam mengatasi masalah imbalanced dataset dengan SMOTE mampu meningkatkan akurasi SVM sebesar 2% hingga 23%. Peneliti [3] menggunakan ensemble learning untuk melakukan klasifikasi cuaca dengan efektif dengan metode yang digunakan yaitu Random Forest dengan teknik oversampling dimana digunakan untuk menyelesaikan ketidakseimbangan jumlah data dari masing-masing kelas cuaca, yang memiliki kategori cuaca Cerah, Cerah Berawan, Berawan, Berawan Tebal, Hujan Lokal, Hujan Ringan, Hujan Sedang , dan Hujan Petir. Hasil pengujian dengan model Random Forest mencapai tingkat akurasi sebesar 70%. Teknik oversampling yang diterapkan adalah metode Synthetic Minority Over-sampling Technique (SMOTE), yang mampu meningkatkan prediksi dari setiap kelas minoritas secara signifikan dengan rata-rata peningkatan sebesar 50% [3].

Berdasarkan uraian penelitian yang sudah dilakukan, sehingga pada penelitian ini dalam halnya menangani kualitas udara dengan fokus pada peningkatan akurasi klasifikasi kualitas udara dengan menerapkan teknik oversampling pada metode Support Vector Machine (SVM) dan Random Forest. Teknik oversampling diperlukan utuk mengatasi masalah ketidakseimbangan data, dimana kelas minoritas dalam dataset memiliki representasi yang sangat kecil dibandingkan kela mayoritas. Tanpa teknik oversampling, model machine learning seperti SVM dan Random Forest akan mengalami kesulitan dalam mengenali pola penting yang terdapat pada kelas minoritas, yang menyebabkan prediksi tidak akurat. Oversampling membantu dengan menambah jumlah sampel pada kelas minoritas sehingga model dapat belajar dari data yang lebih representatif. Teknik ini dapat meningkatkan kemampuan model untuk mengenali dan mengklasifikasikan pola penting dari data yang minim. Tanpa adanya teknik oversampling model dapat bias terhadap kelas mayoritas dan tidak dapat mengklasifikasi kelas minoritas dengan benar. Hal tersebut dapat menghasilkan rendahnya akurasi dan performa klasifikasi, terutama dalam konteks yang memerlukan pengenalan pola minoritas yang krusial. Dengan menggunakan teknik oversamping pada SVM dan Random Forest, diharapkan dapat meningkatkan akurasi dan keandalan prediksi dalam klasifikasi kualitas udara, dan memberikan manfaat yang signifikan dalam pemantauan dan pengelolaan lingkungan.

2. Metode Penelitian

2.1. Gambaran Umum Sistem

Pada Gambar 1, terdapat gambaran umum sistem untuk membuat model klasifikasi kondisi kualitas udara dengan menggunakan machine learning. Proses awal penelitian dimulai dengan pengumpulan data yang dicari dari pihak terkait yang memiliki wewenang untuk melakukan pengukuran kondisi udara di Jakarta. Data yang diperoleh kemudian diproses dengan tahap preprocessing untuk menyeimbangkan data menggunakan teknik oversampling. Setelah proses oversampling selesai, dataset baru yang dihasilkan digunakan untuk melatih model. Proses pelatihan model dibedakan berdasarkan data yang telah di-oversampling dan data original. Metode machine learning yang digunakan untuk pemodelan adalah Random Forest dan Support Vector Machine. Selanjutnya, dilakukan proses evaluasi model berdasarkan hasil tabel confusion matrix untuk menilai akurasi model. Hasil akurasi pemodelan dibandingkan antara dataset dengan oversampling dan dataset orisinal. Hal ini bertujuan untuk mengetahui peningkatan akurasi model setelah menerapkan teknik oversampling pada data yang tidak seimbang.



Gambar 1. Gambaran Umum Sistem.

2.2. Informasi Dataset

Penelitian ini menggunakan data yang diperoleh dari portal Satu Data yang diterbitkan oleh Dinas Lingkungan Hidup DKI Jakarta. Data Indeks Standar Pencemaran Udara (ISPU) ini terdiri dari 365 baris data, mencakup rentang waktu dari awal tahun 2022 hingga akhir tahun 2022 [4]. Data ini menggambarkan kondisi kualitas udara di DKI Jakarta dengan 13 parameter pengukuran, seperti Periode, Bulan, Tanggal, Stasiun, PM_10, PM_2.5, sulfur_dioksida, karbon_monoksida, ozon, nitrogen_dioksida, max, parameter_pencemar_kritis, dan kategori. Untuk informasi lebih detail, dapat merujuk ke Tabel 1.

Dalam penelitian ini, parameter kategori dijadikan sebagai target klasifikasi. Parameter kategori memiliki tiga jenis klasifikasi untuk menilai kualitas udara, yaitu Baik, Sedang, dan Tidak Sehat. Sedangkan untuk pembuatan model, hanya menggunakan 7 parameter, yaitu PM_10, PM_2.5, sulfur_dioksida, karbon_monoksida, ozon, nitrogen_dioksida, dan kategori.

Tabel 1. Atribut Penyusun Dataset.

Atribut	Deskripsi
Periode	Periode dalam sebulan
Bulan	Bulan pengambilan data
Tanggal	Tanggal pengambilan data
Stasiun	Lokasi pengambilan data
PM_10	Nilai pengukuran untuk parameter PM 10 Mikron
PM_2.5	Nilai pengukuran untuk parameter PM 2.5 Mikron
Sulfur_dioksida	Nilai pengukuran untuk parameter sulfur dioksida (SO2)
Karbon_monoksida	Nilai pengukuran untuk parameter karbon monoksida (CO)
Ozon	Nilai pengukuran untuk parameter ozon (O3)
Nitrogen_dioksida	Nilai pengukuran untuk parameter nitrogen dioksida (NO2)
Max	Nilai tertinggi hasil pengukuran dari beberapa parameter
Parameter_pencemar_kritis	Nama parameter dengan nilai tertinggi
Kategori	Kategori hasil pengukuran (Baik, Sedang dan Tidak Sehat)

2.3. Synthetic Minority Oversampling Technique (SMOTE)

Berdasarkan dataset yang ada, terlihat bahwa jumlah antara kelasnya tidak seimbang. Data tersebut terdiri dari tiga kelas yaitu Baik, Sedang, dan Tidak Sehat. Jumlah data untuk kelas Baik adalah 3, kelas Sedang adalah 225, dan kelas Tidak Sehat adalah 137, sebagaimana yang diperlihatkan pada Tabel 2 (data original) dan Gambar 2a. Penggunaan data yang tidak seimbang seperti ini dapat mempengaruhi akurasi hasil model menjadi rendah. Oleh karena itu, diperlukan proses oversampling untuk menghasilkan data sintetis pada kelas yang memiliki jumlah data sedikit.

Dalam penelitian ini, proses oversampling dilakukan menggunakan teknik SMOTE [5][6]. Teknik ini bekerja dengan menghitung jarak antara fitur yang dipilih dengan fitur tetangga terdekatnya menggunakan metode Euclidean Distance (persamaan 1) [7]. Setelah mendapatkan titik terdekat, dilakukan perhitungan untuk menciptakan data sintetis (persamaan 2) [8]. Proses perhitungan ini juga melibatkan nilai acak antara 0 dan 1 yang dikalikan dengan nilai baru yang dihasilkan, kemudian ditambahkan dengan nilai yang sebelumnya dipilih. Jumlah data pada dataset yang telah di oversampling dengan SMOTE ditampilkan dalam Tabel 2 dan Gambar 2b, sementara sebaran data diperlihatkan dalam Gambar 3.

$$d = \sqrt{(y_1 - x_1)^2 + \cdots + (y_n - x_n)^2} \quad (1)$$

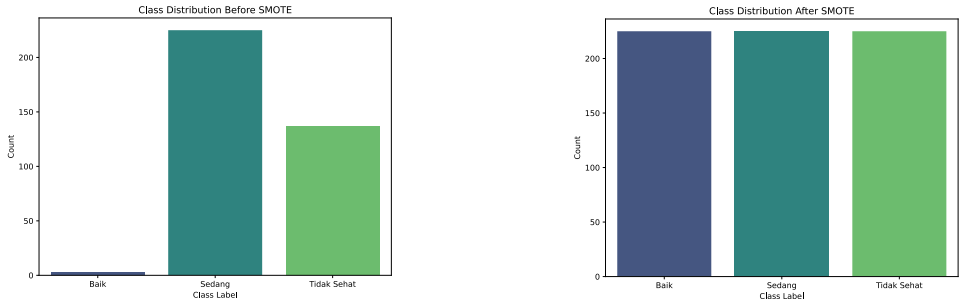
d : jarak antara objek, untuk mendapatkan nilai terdekat dengan titik sample.
y : titik data sample
x : titik perbandingan dengan sample

$$X_{new} = P + rand(0,1) \times (d - P) \quad (2)$$

X_{new} : nilai baru hasil SMOTE
P : titik data sample
d : titik terdekat dengan data sample

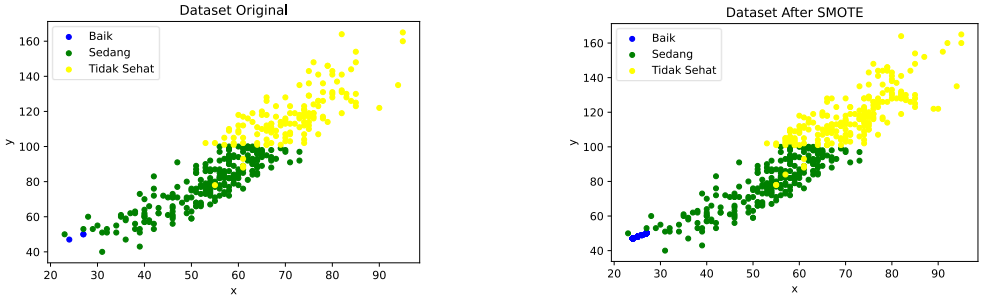
Tabel 2. Informasi Dataset Original dan Setelah Oversampling SMOTE.

No.	Kualitas Udara	Original	Persentase (%)	SMOTE	Persentase (%)
1	Baik	3	0.82	225	33.33
2	Sedang	225	61.64	225	33.33
3	Tidak Sehat	137	37.54	225	33.33
	Total	365	100	675	100



(a) (b)

Gambar 2. Perbandingan Jumlah Data (a) Sebelum SMOTE, (b) Setelah SMOTE.



(a) (b)

Gambar 3. Sebaran Data (a) Sebelum SMOTE, (b) Setelah SMOTE.

2.4. Random Forest

Random Forest adalah metode dalam Machine Learning yang menggabungkan teknik-teknik dari pohon keputusan dan bagging untuk melakukan prediksi [9]. Proses ini melibatkan pembagian data menjadi beberapa cabang sampai kriteria berhenti terpenuhi, kemudian prediksi dilakukan dengan berpindah melalui simpul dan cabang. Langkah-langkahnya dimulai dengan menentukan fitur yang menjadi akar, diikuti dengan penghitungan nilai entropi sebagai tahap pertama. Setelah entropi diperoleh, langkah selanjutnya adalah menghitung gain informasi. Dari hasil gain informasi pada data, dilakukan generalisasi terhadap pohon keputusan. Secara keseluruhan, persamaan untuk entropi dan gain informasi umumnya ditunjukkan pada persamaan 2 dan 3.

$$Entrophy\ (S) \sum_{i=1}^{n} P_i\ log2P_i \tag{2}$$

$$Gain\ (S,A) = E(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} E(S_i) \tag{3}$$

2.5. Support Vector Machine

Algoritma Support Vector Machine (SVM) yaitu algoritma yang berfokus dalam klasifikasi pada data linear dan nonlinear [10]. Pada Tabel 3 memperlihatkan klasifikasi data linear dan non linear dengan jenis kernel dan persamaan rumusnya. Dalam penelitian ini untuk melakukan prediksi kualitas udara dengan SVM menggunakan kernel RBF.

Tabel 3. SVM Kernel.

SVM	Jenis Kernel	Persamaan
Linier	Linier	$K\ (x,y) = x.y$
Non Linier	Polynomial	$K\ (x,y) = (x.y+1)^p$
	RBF	$K\ (x,y) = e^{-\lvert x.y\rvert^2_{\square}/2\sigma^2}$
	Sigmoid	$(x,y) = tanh(Kx.y - \delta)$

2.6. Evaluasi Model

Proses evaluasi model menggunakan teknik confusion matrix, yang merupakan tabel dengan empat kombinasi berbeda dari nilai prediksi dan nilai aktual. Keempat kombinasi menunjukkan representasi hasil proses klasifikasi pada confusion matrix, yaitu [11]:

a. True Positive (TP) adalah data positif yang diprediksi dengan benar.
b. True Negative (TN) adalah data negatif yang diprediksi dengan benar.
c. False Positive (FP) adalah data negatif yang salah diprediksi sebagai positif.
d. False Negative (FN) adalah data positif yang salah diprediksi sebagai negatif.

Pada umumnya, confusion matrix berisi informasi yang membandingkan hasil klasifikasi yang diberikan oleh suatu sistem dengan hasil klasifikasi yang seharusnya. Melalui nilai-nilai yang terdapat dalam tabel confusion matrix tersebut, akan dapat digunakan untuk menghitung nilai akurasi, presisi, recall, dan F1-measure. Perhitungan untuk nilai ini diperlihatkan pada Tabel 4.

Tabel 4. Matrix Evaluasi.

Matrix	Persamaan
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$
Precision	$\frac{TP}{TP+FP}$
Recall	$\frac{TP}{TP+FN}$
F-Measure	$F_\beta = \frac{1}{\beta \times \frac{1}{precision} + (1-\beta) \times \frac{1}{recall}}$

3. Hasil dan Pembahasan

Pada bagian ini, hasil dari proses pengujian yang telah dilakukan dijelaskan. Hasil pengujian dibedakan menjadi empat model, yaitu:

a. Pemodelan dengan menggunakan metode Random Forest dengan data original.
b. Pemodelan dengan menggunakan metode Random Forest dengan data yang telah di oversampling SMOTE.
c. Pemodelan dengan menggunakan metode Support Vector Machine dengan data original.
d. Pemodelan dengan menggunakan metode Support Vector Machine dengan data yang telah di oversampling SMOTE.

Kemudian, berdasarkan hasil yang diperoleh dari tabel confusion matrix untuk masing-masing pemodelan dijelaskan sebagai berikut, pada Tabel 5, terdapat hasil confusion matrix untuk metode Random Forest dengan data original. Nilai *true positive* untuk kelas Baik adalah 5, Sedang adalah 224, dan Tidak Sehat adalah 134. Selanjutnya, untuk metode Random Forest dengan data SMOTE, didapatkan confusion matrix yang diperlihatkan pada Tabel 6. Nilai true positive untuk kelas Baik adalah 225, Sedang adalah 223, dan Tidak Sehat adalah 223.

Berikutnya, untuk pengujian dengan metode Support Vector Machine menggunakan data original, didapatkan tabel confusion matrix seperti yang diperlihatkan pada Tabel 7. Nilai true positive untuk kondisi Baik adalah 0, Sedang adalah 215, dan Tidak Sehat adalah 118. Sedangkan, jika menggunakan data SMOTE dengan metode Support Vector Machine, didapatkan nilai confusion matrix seperti yang diperlihatkan pada Tabel 8. Untuk kondisi Baik, jumlahnya adalah 225, Sedang adalah 195, dan Tidak Sehat adalah 219.

Tabel 5. Confusion Matrix untuk Random Forest.

	Baik	Sedang	Tidak Sehat
Baik	2	1	0
Sedang	1	224	0
Tidak Sehat	0	3	134

Tabel 6. Confusion Matrix untuk Random Forest dengan Oversampling SMOTE.

	Baik	Sedang	Tidak Sehat
Baik	225	0	0
Sedang	1	223	1
Tidak Sehat	0	2	223

Tabel 7. Confusion Matrix untuk Support Vector Machine.

	Baik	Sedang	Tidak Sehat
Baik	0	3	0
Sedang	0	215	10
Tidak Sehat	0	19	118

Tabel 8. Confusion Matrix untuk Support Vector Machine dengan Oversampling SMOTE.

	Baik	**Sedang**	**Tidak Sehat**
Baik	225	0	0
Sedang	13	195	17
Tidak Sehat	0	6	219

Setelah mendapatkan nilai confusion matrix dari masing-masing pemodelan, berikutnya dilakukan perhitungan untuk mencari nilai akurasi, recall, precision, dan F1-measure. Pada Tabel 9 memperlihatkan hasil nilai akurasi untuk setiap pemodelan. Pemodelan dengan Random Forest mendapatkan akurasi sebesar 98%, sedangkan pemodelan Random Forest dengan SMOTE mengalami peningkatan akurasi menjadi 99%. Pada pemodelan dengan Support Vector Machine, akurasi yang diperoleh adalah 91%, sementara pemodelan Support Vector Machine dengan SMOTE mengalami peningkatan akurasi menjadi 95%. Selain itu, hasil perhitungan untuk nilai recall diperlihatkan pada Tabel 10. Hasil perhitungan untuk nilai precision diperlihatkan pada Tabel 11, sedangkan hasil perhitungan untuk nilai F1-measure diperlihatkan pada Tabel 12.

Tabel 9. Hasil Akurasi.

Metode	**Accuracy (%)**
Random Forest	98
Random Forest + SMOTE	99
Support Vector Machine	91
Support Vector Machine + SMOTE	95

Tabel 10. Hasil Recall.

Method	**Recall Baik (%)**	**Recall Sedang (%)**	**Recall Tidak Sehat (%)**
Random Forest	67	100	98
Random Forest dan SMOTE	100	99	99
Support Vector Machine	0	96	86
Support Vector Machine + SMOTE	100	87	97

Tabel 11. Hasil Precision.

Method	**Precision Baik (%)**	**Precision Sedang (%)**	**Precision Tidak Sehat (%)**
Random Forest	67	98	100
Random Forest dan SMOTE	100	99	100
Support Vector Machine	0	91	92
Support Vector Machine + SMOTE	95	97	93

Tabel 12. Hasil F1-Measure.

Method	F1-Measure Baik (%)	F1-Measure Sedang (%)	F1-Measure Tidak Sehat (%)
Random Forest	67	99	99
Random Forest dan SMOTE	100	99	99
Support Vector Machine	0	93	89
Support Vector Machine + SMOTE	97	92	95

4. Kesimpulan

Penelitian ini bertujuan untuk membuat model yang dapat mengklasifikasi kondisi kualitas udara. Data yang digunakan adalah data Indeks Standar Pencemaran Udara (ISPU) selama tahun 2022 yang diterbitkan oleh Dinas Lingkungan Hidup DKI Jakarta. Data ini terdiri dari 13 parameter pengukuran kualitas udara, di mana 7 parameter di antaranya digunakan untuk membuat model dalam penelitian. Satu parameter digunakan sebagai target, yaitu parameter kategori kualitas udara (Baik, Sedang, Tidak Sehat). Karena jumlah data untuk setiap kategori tidak seimbang, dilakukan oversampling menggunakan teknik SMOTE. Setelah proses oversampling, jumlah data menjadi seimbang dengan masing-masing target sebanyak 225, sehingga total data menjadi 675.

Selanjutnya, dilakukan proses pemodelan menggunakan Metode Random Forest dan Support Vector Machine. Hasil pemodelan dibandingkan antara data oversampling SMOTE dan data tanpa oversampling (data orisinal). Hasilnya pemodelan dengan Random Forest mendapatkan akurasi sebesar 98%, sedangkan pemodelan Random Forest dengan SMOTE mengalami peningkatan akurasi menjadi 99%. Pada pemodelan dengan Support Vector Machine, akurasi yang diperoleh adalah 91%, sementara pemodelan Support Vector Machine dengan SMOTE mengalami peningkatan akurasi menjadi 95%. Dari hasil tersebut, dapat disimpulkan bahwa setiap model dengan metode Random Forest ataupun Support Vector Machine memiliki hasil akurasi yang meningkat saat menggunakan data yang telah di-oversampling dengan SMOTE dibandingkan dengan data yang tidak seimbang atau tanpa SMOTE.

Daftar Pustaka

[1] A. S. Handayani, S. Soim, T. E. Agusdi, Rumiasih, and A. Nurdin, "Klasifikasi Kualitas Udara Dengan Metode Support Vector Machine," *JIRE (Jurnal Informatika & Rekayasa Elektronika)*, vol. 3, no. 2, pp. 187–199, 2020.

[2] A. Indrawati, "Penerapan Teknik Kombinasi Oversampling Dan Undersampling Untuk Mengatasi Permasalahan Imbalanced Dataset," *Jurnal Informatika dan Komputer) Akreditasi KEMENRISTEKDIKTI*, vol. 4, no. 1, 2021, doi: 10.33387/jiko.

[3] F. Hamami and I. A. Dahlan, "Klasifikasi Cuaca Provinsi Dki Jakarta Menggunakan Algoritma Random Forest Dengan Teknik Oversampling," *Jurnal Teknoinfo*, vol. 16, no. 1, p. 87, 2022, doi: 10.33365/jti.v16i1.1533.

[4] D. L. Hidup, "Data Indeks Standar Pencemar Udara (ISPU) di Provinsi DKI Jakarta," 2022. Accessed: Jan. 01, 2022. [Online]. Available: https://satudata.jakarta.go.id

[5] G. A. Pradipta, R. Wardoyo, A. Musdholifah, I. N. H. Sanjaya, and M. Ismail, "SMOTE for Handling Imbalanced Data Problem : A Review," in *2021 Sixth International Conference on Informatics and Computing (ICIC)*, 2021, pp. 1–8. doi: 10.1109/ICIC54025.2021.9632912.

[6] R. Das, S. Kr. Biswas, D. Devi, and B. Sarma, "An Oversampling Technique by Integrating Reverse Nearest Neighbor in SMOTE: Reverse-SMOTE," in *2020 International Conference on Smart Electronics and Communication (ICOSEC)*, 2020, pp. 1239–1244. doi: 10.1109/ICOSEC49089.2020.9215387.

[7] G. Putro Dirgantoro, M. A. Soeleman, and C. Supriyanto, "Smoothing Weight Distance to Solve Euclidean Distance Measurement Problems in K-Nearest Neighbor Algorithm," in *2021 IEEE 5th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, 2021, pp. 294–298. doi: 10.1109/ICITISEE53823.2021.9655820.

[8] R. Wardoyo, I. M. A. Wirawan, and I. G. A. Pradipta, "Oversampling Approach Using Radius-SMOTE for Imbalance Electroencephalography Datasets," *Emerging Science Journal*, vol. 6, no. 2, pp. 382–398, 2022, doi: 10.28991/ESJ-2022-06-02-013.

[9] A. L. Latifah, A. Shabrina, I. N. Wahyuni, and R. Sadikin, "Evaluation of Random Forest model for forest fire prediction based on climatology over Borneo," in *2019 International Conference on Computer, Control, Informatics and its Applications (IC3INA)*, 2019, pp. 4–8. doi: 10.1109/IC3INA48034.2019.8949588.

[10] Y. Dong, H. Wang, L. Zhang, and K. Zhang, "An improved model for PM2.5 inference based on support vector machine," in *2016 17th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)*, 2016, pp. 27–31. doi: 10.1109/SNPD.2016.7515873.

[11] D. Normawati and S. A. Prayogi, "Implementasi Naïve Bayes Classifier Dan Confusion Matrix Pada Analisis Sentimen Berbasis Teks Pada Twitter," *Jurnal Sains Komputer & Informatika (J-SAKTI*, vol. 5, no. 2, pp. 697–711, 2021.